Building Interoperable Digital Libraries: A Practical Guide to creating Open Archives

Hussein Suleman, hussein@vt.edu
Digital Library Research Laboratory
Virginia Tech
1. Introduction

- What is the OAI?
- Motivation
- General System Strategy
- History
- Case study: NDLTD
1.1. What is the OAI?

What is the Open Archives Initiative (OAI)?
- Organization dedicated to solving problems of digital library interoperability by defining simple protocols, most recently for the exchange of metadata.

What is the Protocol for Metadata Harvesting?
- Protocol to transfer metadata from a source archive to a destination archive.
1.2. Motivation

- Existence of some established but independent archives
- Need for cross-archive services (like search engines)
- Lack of low-cost interoperability technology
- Experience from past projects such as Dienst
1.3. General System Strategy

- Services
- Metadata Harvesting
- Document Model
1.4. History

Santa Fe Meeting – October 1999
- Santa Fe Convention, January 2000

Workshops (ACM-DL 2000, ECDL 2000)

Structuring of the OAI
- Steering Committee
- Technical Committee

Open Days – US/Europe
- Protocol for Metadata Harvesting v1.0, January 2001

Minor Update: v1.1 – July 2001

Version 2.0 – June 2002
1.5. Case Study: NDLTD

Networked Digital Library of Theses and Dissertations

Multiple independent university-based collections of electronic documents

- Virginia Tech
- Humboldt U.
- U. South Florida

OAI Protocol for Metadata Harvesting

International ETD Library
2. Definitions / Concepts

Basic Principles
- What is an Open Archive?
- Harvesting vs. Federation
- Metadata vs. Data
- Data and Service Providers

Underlying Technology
- HTTP and XML
- XML, XML Namespaces and Schema

Protocol Policies
- Uniqueness and Persistence
- What is a record?
- Multiplicity of Metadata
- Sets
- Datestamp, Harvesting and Flow Control
2.1. What is an Open Archive?

- Any WWW-based system that can be accessed through the well-defined interface of the Open Archives Protocol for Metadata Harvesting
- ... a.k.a. OAI-Compliant Repository
- No implications for:
  - Physical storage of data
  - Cost of data
  - Metadata and data formats
  - Access control to server
2.2. Harvesting vs. Federation

- Competing approaches to interoperability
  - Federation is when services are run remotely on remote data (e.g. Federated searching)
  - Harvesting is when data/metadata is transferred from the remote source to the destination where the services are located (e.g. Union catalogues)
  - Federation requires more effort at each remote source but is easier for the local system and vice versa for harvesting
  - OAI currently focuses on harvesting
2.3. Metadata vs. Data

- Data refers to digital objects or digital representations of objects.
- Metadata is information about the objects (e.g. title, author, etc.).
- OAI focuses on metadata, with the implicit understanding that metadata usually contains useful links to the source digital objects.
2.4. Data and Service Providers

Data Providers refer to entities who possess data/metadata and are willing to share this with others (internally or externally) via well-defined OAI protocols (e.g. database servers).

Service Providers are entities who harvest data from Data Providers in order to provide higher-level services to users (e.g. search engines).

OAI uses these denotations for its client/server model (data=server, service=client).
2.5. HTTP and XML

- Metadata Harvesting Protocol is an almost stateless request/response protocol
- Requests and responses are sent via the HTTP protocol
- Requests are encoded as GET/POST operations
- Responses are well-formed XML documents
2.6. XML Namespaces and Schema

Consistency and data quality is ensured by using XML Schema descriptions for each possible response.

XML Namespaces are used where necessary to clearly define which parts of the responses are actual metadata and which support the Metadata Harvesting Protocol.
2.7. Uniqueness and Persistence

- Each record must be uniquely addressable by a distinct identifier.
- Each metadata entity should ideally be persistent to guarantee that service providers can always refer back to the source.
2.8. What is a record?

- A record refers to an independent XML structure that may be associated with digital or physical objects.
- Records are usually associated with metadata, not data.
- OAI advocates harvesting of records, which contain metadata and additional fields to support the harvesting operation.
2.9. Sample OAI Record

(note: schema and namespaces have been left out for clarity)

<record>
  <header>
    <identifier>oai:jcdl2002.org:tut1</identifier>
    <datestamp>2002-02-03</datestamp>
    <setSpec>tut</setSpec>
  </header>
  <metadata>
    <dc>
      <title>OAI Tutorial at JCDL</title>
      <creator>Hussein Suleman</creator>
      <language>English</language>
    </dc>
  </metadata>
  <about>
    <metadataID>oai:jcdl2002.org:tut1md</metadataID>
  </about>
</record>
2.10. Multiplicity of Metadata

- Multiple formats of metadata allowed
- Dublin Core is mandatory
- Any other format allowed as long as it has an XML encoding
- E.g. MARC (Libraries), IMS (Education), ETDMS (Theses/Dissertations), RFC1807 (Bibliographies)
2.11. Sets

- Protocol mechanism to allow for harvesting of sub-collections
- No well-defined semantics – depends completely on local data providers
- May be defined by arrangement between data providers and service providers
- E.g. Subject areas, years, author names, search queries
2.12. Datestamps & Harvesting

Each record needs a datestamp that indicates its date of creation or modification.

Dates are used to allow for harvesting by date range, thus allowing incremental and continuous transfer of metadata from a data provider to a service provider.
2.13. Flow Control

- HTTP “retry-after” mechanism can be leveraged to support server-side delaying of a client’s request.
- Resumption Tokens can be used to return partial results – the client is issued with a token which may be presented to the server to receive more results.
3. Requirements to be a Data Provider

- Source of metadata
- Server technology
- Datestamps
- Deletions
- Unique identifiers
- Metadata mappings
3.1. Source of Metadata

- Database in proprietary format
- Collection of metadata records in well-defined format/s
  - Files on disk
- Metadata may be dynamically or statically extracted from data
- Synthetic collection
3.2. Server Technology

- WWW Server
- Protocol may be implemented in many forms
  - CGI Script (Perl, C++, Java)
  - Java Servlet
  - PHP
- Metadata (e.g. database) access mechanism required
- See www.openarchives.org for list of publicly available software templates
- See www.dlib.vt.edu for VT experimental software
3.3. Datestamps

- Needed for every record to support incremental harvesting
- Must be updated for every addition/modification/deletion to ensure changes are correctly propagated
- Different from dates within the metadata – this date is used only for harvesting
- Can be either YYYY-MM-DD or YYYY-MM-DDThh:mm:ssZ (must be GMT timezone)
3.4. Unique Identifiers

- Each record must have a unique identifier
- Identifiers must be valid URIs
- Example:
  - oai:<archiveId>;<recordId>
  - oai:etd.vt.edu:etd-1234567890
- Each identifier must resolve to a single record and always to the same record (for a given metadata format)
3.5. Deletions

- Archives may keep track of deleted records, by identifier and datestamp
- All protocol result sets can indicate deleted records
- If deletions are being tracked, this information must be stored indefinitely so as to correctly propagate to service providers with varying harvesting schedules
3.6. Metadata Mappings

- Data provider must map its metadata to the formats it chooses to provide through its OAI interface
- Unqualified Dublin Core required
  - Best practice is to include a link in the `<identifier>` tag to the actual digital resource or at least a human-readable web page
- Native formats recommended
- Community-based formats recommended
4. Metadata Harvesting Protocol

- Service Requests
  - Identify
  - ListMetadataFormats
  - ListSets
  - GetRecord
  - ListIdentifiers
  - ListRecords

- Metadata Multiplicity
- Date Ranges
- Resumption Tokens
- Error and Exceptions
4.1. Identify

- **Purpose**
  - Return general information about the archive and its policies

- **Parameters**
  - None

- **Sample URL**
  - http://www.anarchive.org/cgi-bin/OAI?verb=Identify
4.2. Identify - Response

```xml
<?xml version="1.0" encoding="UTF-8" ?>
<OAI-PMH xmlns="http://www.openarchives.org/OAI/2.0/">
  xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
  xsi:schemaLocation="http://www.openarchives.org/OAI/2.0/
http://www.openarchives.org/OAI/2.0/OAI-PMH.xsd">
<responseDate>2002-05-26T19:28:31Z</responseDate>
:request verb="Identify">http://rocky.dlib.vt.edu/~jcdlpix/cgi-bin/OAI2.0/beta2/jcdl/oai.pl</request>
  <Identify>
    <repositoryName>JCDL 2001 Picture Archive</repositoryName>
    <baseURL>http://rocky.dlib.vt.edu/~jcdlpix/cgi-bin/OAI2.0/beta2/jcdl/oai.pl</baseURL>
    <protocolVersion>2.0b2</protocolVersion>
    <adminEmail>jcdlpix@rocky.dlib.vt.edu</adminEmail>
    <earliestDatestamp>1970-01-01T00:00:00Z</earliestDatestamp>
    <deletedRecord>no</deletedRecord>
    <granularity>YYYY-MM-DD</granularity>
  </Identify>
</OAI-PMH>
```
4.3. ListMetadataFormats

Purpose

- List metadata formats supported by the archive as well as their schema locations and namespaces

Parameters

- identifier – for a specific record (O)

Sample URL

4.4. ListMetadataFormats - Response

```xml
<?xml version="1.0" encoding="UTF-8" ?>
<OAI-PMH xmlns="http://www.openarchives.org/OAI/2.0/">
  xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
  xsi:schemaLocation="http://www.openarchives.org/OAI/2.0/
  http://www.openarchives.org/OAI/2.0/OAI-PMH.xsd">
  <responseDate>2002-05-26T19:27:25Z</responseDate>
  <request verb="ListMetadataFormats">http://rocky.dlib.vt.edu/~jcdlpx/cgi-bin/OAI2.0/beta2/jcdl/oai.pl</request>
  <ListMetadataFormats>
    <metadataFormat>
      <metadataPrefix>oai_dc</metadataPrefix>
      <schema>http://www.openarchives.org/OAI/2.0/oai_dc.xsd</schema>
      <metadataNamespace>http://www.openarchives.org/OAI/2.0/oai_dc/</metadataNamespace>
    </metadataFormat>
  </ListMetadataFormats>
</OAI-PMH>
```
4.5. ListSets

**Purpose**
- Provide a hierarchical listing of sets in which records may be organized

**Parameters**
- None

**Sample URL**
4.6. ListSets – Response

```xml
<?xml version="1.0" encoding="UTF-8" ?>
<OAI-PMH xmlns="http://www.openarchives.org/OAI/2.0/
   xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
   xsi:schemaLocation="http://www.openarchives.org/OAI/2.0/
   http://www.openarchives.org/OAI/2.0/OAI-PMH.xsd">
  <responseDate>2002-05-26T19:29:52Z</responseDate>
  <request verb="ListSets">http://rocky.dlib.vt.edu/~jcdlpx/cgi-bin/OAI2.0/beta2/jcdl/oai.pl</request>
  <ListSets>
    <set>
      <setSpec>200105dle</setSpec>
      <setName>JCDL Day Four</setName>
      <setDescription>
        <oaidc:dc
          xmlns:oaidc="http://www.openarchives.org/OAI/2.0/oai_dc/"
          xmlns="http://purl.org/dc/elements/1.1/"
          xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
          xsi:schemaLocation="http://www.openarchives.org/OAI/2.0/oai_dc/
             http://www.openarchives.org/OAI/2.0/oai_dc.xsd">
          <description>Pictures taken during JCDL Day Four</description>
        </oaidc:dc>
      </setDescription>
    </set>
  </ListSets>
</OAI-PMH>
```
4.7. GetRecord

**Purpose**

- Returns the metadata for a single identifier in the form of an OAI record

**Parameters**

- identifier – unique id for record (R)
- metadataPrefix – metadata format (R)

**Sample URL**

4.8. GetRecord - Response

```
<responseDate>2002-05-26T19:32:54Z</responseDate>
:request verb="GetRecord" metadataPrefix="oai_dc"
    identifier="oai:JCDLPICS:200105dle1">http://rocky.lib.vt.edu/~jcdlpx/cgi-bin/OAI2.0/beta2/jcdl/oai.pl</request>
  - <GetRecord>
    - <record>
      - <header>
        <identifier{oai:JCDLPICS:200105dle1}</identifier>
        <datestamp>2001-06-27</datestamp>
        <setSpec>200105dle</setSpec>
      </header>
      - <metadata>
        - <oai:dc xmlns="http://purl.org/dc/elements/1.1/"
          xmlns:oai="http://www.openarchives.org/OAI/2.0/oai_dc/
          xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
          xsi:schemaLocation="http://www.openarchives.org/OAI/2.0/oai_dc/
          http://www.openarchives.org/OAI/2.0/oai_dc.xsd">
          <title>01dle1</title>
          <creator>Hussein Suleman</creator>
          <subject>JCDL Day Four</subject>
          <description>Jim French and Unmil Karadkar over lunch</description>
          <publisher>JCDL</publisher>
          <date>2001-06-27</date>
          <type>image</type>
```

JCDL 2002
4.9. ListIdentifiers

Purpose

- List headers for all records corresponding to the specified parameters

Parameters

- from – start date (O)
- until – end date (O)
- set – set to harvest from (O)
- metadataPrefix – metadata format to list identifiers for (R)
- resumptionToken – flow control mechanism (X)

Sample URL

4.10. ListIdentifiers - Response

```xml
<?xml version="1.0" encoding="UTF-8" ?>
<OAI-PMH xmlns="http://www.openarchives.org/OAI/2.0/
xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
xsi:schemaLocation="http://www.openarchives.org/OAI/2.0/
http://www.openarchives.org/OAI/2.0/OAI-PMH.xsd">
  <responseDate>2002-05-26T19:36:13Z</responseDate>
  <request verb="ListIdentifiers">
    <metadataPrefix>oai_dc</metadataPrefix>
    http://rocky.dlib.vt.edu/~jcdlpix/cgi-bin/OAI2.0/beta2/jcdl/oai.pl</request>
  - <ListIdentifiers>
    - <header>
      <identifier>oai:JCDLPICS:200105dle1</identifier>
      <datestamp>2001-06-27</datestamp>
      <setSpec>200105dle</setSpec>
    </header>
    - <header>
      <identifier>oai:JCDLPICS:200105dle2</identifier>
      <datestamp>2001-06-27</datestamp>
      <setSpec>200105dle</setSpec>
    </header>
</OAI-PMH>
```
4.11. ListRecords

**Purpose**
- Retrieves metadata for multiple records

**Parameters**
- from – start date (O)
- until – end date (O)
- set – set to harvest from (O)
- resumptionToken – flow control mechanism (X)
- metadataPrefix – metadata format (R)

**Sample URL**
- http://www.anarchive.org/cgi-bin/OAI?
  verb=ListRecord&metadataPrefix=oai_dc&from=2001-01-01
4.12. ListRecords - Response

```xml
<Address>du/~jcdlpix/cgi-bin/OAI2.0/beta2/jcdl/oai.pl?verb=ListRecords&metadataPrefix=oai_dc</Address>
<xsi:schemaLocation="http://www.openarchives.org/OAI/2.0/
http://www.openarchives.org/OAI/2.0/OAI-PMH.xsd">
<responseDate>2002-05-26T19:37:37Z</responseDate>
<request verb="ListRecords"
    metadataPrefix="oai_dc">http://rocky.dlib.vt.edu/~jcdlpix/cgi-bin/OAI2.0/beta2/jcdl/oai.pl</request>
- <ListRecords>
- <record>
    - <header>
        <identifier>oai:JCDLPICS:200105dle1</identifier>
        <datestamp>2001-06-27</datestamp>
        <setSpec>200105dle</setSpec>
    </header>
    - <metadata>
        - <oaidc:dc xmlns="http://purl.org/dc/elements/1.1/"
            xmlns:oaidc="http://www.openarchives.org/OAI/2.0/oai_dc/"
            xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
            xsi:schemaLocation="http://www.openarchives.org/OAI/2.0/oai_dc/
http://www.openarchives.org/OAI/2.0/oai_dc.xsd">
            <title>01dle</title>
            <creator>Hussein Suleman</creator>
            <subject>JCDL Day Four</subject>
            <description>Jim French and Unmil Karadkar over lunch</description>
        </oaidc:dc>
    </metadata>
</record>
</ListRecords>
```

JCDL 2002
4.13. Metadata Multiplicity

- <record>
  - <header>
    <identifier>oai:VTEDT:etd-3123162539751141</identifier>
    <datestamp>1997-04-22</datestamp>
  </header>
  - <metadata>
http://www.openarchives.org/OAI/rfc1807.xsd">
      <bib-version>1</bib-version>
      <id>etd-3123162539751141</id>
      <entry>1997-04-22</entry>
      <organization>Virginia Polytechnic Institute and State University</organization>
      <title>SMA-Induced Deformations In general Unsymmetric Laminates</title>
      <type>Thesis/Dissertation</type>
    </rfc1807>
  </metadata>
</record>
4.14. Date Ranges

```xml
<?xml version="1.0" encoding="UTF-8" ?>
<OAI-PMH xmlns="http://www.openarchives.org/OAI/2.0/
xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
xsi:schemaLocation="http://www.openarchives.org/OAI/2.0/
http://www.openarchives.org/OAI/2.0/OAI-PMH.xsd">
  <responseDate>2002-05-26T19:41:16Z</responseDate>
  <ListIdentifiers>
    <header>
      <identifier,oai:JCDLPICS:200102dlb1</identifier>
      <datestamp>2001-06-26</datestamp>
      <setSpec>200102dlb</setSpec>
    </header>
    <header>
      <identifier,oai:JCDLPICS:200102dlb2</identifier>
      <datestamp>2001-06-26</datestamp>
      <setSpec>200102dlb</setSpec>
    </header>
  </ListIdentifiers>
</OAI-PMH>
```
4.15. Resumption Token

```xml
<Address \ dl/oai.pl?verb=ListIdentifiers&metadataPrefix=oai_dc&from=2001-06-26&until=2001-06-26

<identifier>oai:JCDLPICS:200101dla9</identifier>
<datestamp>2001-06-26</datestamp>
<setSpec>200101dla</setSpec>
</header>

- <header>
  <identifier>oai:JCDLPICS:200101dla10</identifier>
  <datestamp>2001-06-26</datestamp>
  <setSpec>200101dla</setSpec>
</header>

<resumptionToken cursor="0" completeListSize="35">!2001-06-26!
  2001-06-26!oai_dc!30</resumptionToken>
</ListIdentifiers>
</OAI-PMH>
```
4.16. Errors and Exceptions

```xml
<?xml version="1.0" encoding="UTF-8" ?>
<OAI-PMH xmlns="http://www.openarchives.org/OAI/2.0/
xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
xsi:schemaLocation="http://www.openarchives.org/OAI/2.0/http://www.openarchives.org/OAI/2.0/OAI-PMH.xsd">
<responseDate>2002-05-26T19:43:59Z</responseDate>
<error code="noRecordsMatch">The combination of the values of arguments results in an empty set</error>
</OAI-PMH>
```
5. Implementation Details

- Tools Required
- Basic program layout
- Object-oriented approaches
- Extensible metadata generation
- Data cleaning
- Caching of results
- Error handling
- Denial-of-service prevention
- Creating resumption tokens
5.1. Tools Required

- Code templates if available (available for many languages)
- Basic programming environment
- XML generators (for non-trivial encoding)
- Database access libraries/drivers (e.g. DBI, ODBC, JDBC)
5.2. Basic program layout

parse WWW request to extract parameters
if (verb='Identify')
    ProcessIdentify;
else if (verb='ListMetadataFormats')
    ProcessListMetadataFormats;
else if (verb='ListSets')
    ProcessListSets;
else if (verb='GetRecord')
    ProcessGetRecord;
else if (verb='ListIdentifiers')
    ProcessListIdentifiers;
else if (verb='ListRecords')
    ProcessListRecords;
else
    ReportError ('badVerb');
5.3. Object-Oriented Approaches

- Cleaner separation of protocol, database access and metadata generation

Example approaches

- Each service request is handled by an object
  - Simpler incremental development
- Protocol, Database and Metadata are objects
  - Greater portability of code
- Inheritance from a basic OAI data provider
5.4. Metadata Generation

Approaches

- Map from source to each metadata format
- Use crosswalks (maybe XSLT) to generate additional formats

<table>
<thead>
<tr>
<th>source</th>
<th>dc</th>
<th>rfc1807</th>
</tr>
</thead>
<tbody>
<tr>
<td>name</td>
<td>title</td>
<td>title</td>
</tr>
<tr>
<td>author</td>
<td>creator</td>
<td>author</td>
</tr>
</tbody>
</table>
5.5. Data Cleaning

- Escape special XML characters
- Convert to UTF-8 version of Unicode
- Convert entity references
- Remove extraneous whitespace
- Convert CR/LF for paragraphs
- URLs
  - `/=?#=&;+` must be encoded as escape sequences
5.6. Result Caching

- For multiple requests from many clients or to handle partial result sets
- Keep temporary tables/files
- Expire temporary data when no longer needed
- Is this necessary to handle date-range requests where new items are added to the result set while harvesting is in progress?
5.7. Error Handling

- All protocol errors are in XML format
  - badVerb: illegal verb requested
  - badArgument: illegal parameter values or combinations
  - badResumptionToken, cannotDisseminateFormat, idDoesNotExist: parameters are in right format but are not legal under current conditions
  - noRecordsMatch, noMetadataFormats, noSetHierarchy: empty response exception
5.8. Denial-of-Service Prevention

- Return only partial results and issue a resumption token for more
- Use 503 retry-after HTTP errors to have clients try again after a specified back-off time
- Use access control lists to limit who may access the archive
- Invoke an explicit delay before sending back results
5.9. Creating resumptionTokens

- Combine from/until/metadataPrefix/set and a record number indicator with delimiters into a sequential token
- For example:
  - `from!until!metadataPrefix!set!recordnumber`
  - `2000-01-01!2001-01-01!!All!100`

- Use a session manager with automatic expiry
- For example:
  - `vtetd14june10amsession12`
6. Common Problems

- No unique identifiers!
- No datestamps!
- Incomplete information in database
- New metadata format
- XML responses not validating
6.1. No unique identifiers

- Create an independent identifier mapping
- Use row numbers for a database
- Use filenames for data in files
- Use a hash from other fields
  - E.g. author+year+first word in title
6.2. No datestamps

- Ignore the datestamp parameters and stamp all records with the current date.
- Create a date table with the current date for all old entries and update dates for new entries.
- Most Important: Any harvesting algorithm that is interoperably stable for an archive with real dates should be stable for an archive with synthesized dates.
6.3. Incomplete information

Synthesize metadata fields based on a priori knowledge of the data

Example: publisher and language may be hard-coded for many archives

Omit fields that cannot be filled in correctly – better to have less information than incorrect information!
6.4. New metadata format

Find the description, namespace and formal name of the standard

Find an XML Schema description of the data format

If none exists, write one (consult other OAI people for assistance)

Create the mapping and test that it passes XML schema validation
6.5. XML not validating

- Check namespaces and schema
- Use Repository Explorer in non-validating mode to check structure of XML, without looking at namespaces or schemata
- Validate schema by itself if it is non-standard
- Look at XML produced by other repositories
- Watch out for character encoding issues
7. Tools for Testing

- Repository Explorer
  - Interactive Browsing
  - Testing of parameters
  - Multiple views of data
  - Multilingual support
  - Automatic test suite
- OAI Registry
- XML Schema Validator
7.1. RE Interactive Browsing

Open Archives Initiative - Repository Explorer

This site presents an interface to interactively test archives for compliance with the OAI Protocol for Metadata Harvesting [Click here for details]

JavaScript is required

Note: To avoid HTTP errors, please wait for each page to finish loading before clicking on any link.

Please enter the URL to the OAI interface (everything before the ?) or choose a predefined archive from the table. Then click on a verb from the list below to test that function (entering parameters as necessary)

URL: ___________________________

Humboldt University Berlin, Document Server
JCDL Picture Album
{1.1} A Celebration of Women Writers
{1.1} AISRI (American Indian Studies Research Institute)

[View Archive Website][Test and Add an archive to this list]
7.2. RE Parameter Testing

<table>
<thead>
<tr>
<th>Verbs</th>
<th>Parameters</th>
</tr>
</thead>
<tbody>
<tr>
<td>Identify</td>
<td>from (eg., YYYY-MM-DD):</td>
</tr>
<tr>
<td>List Metadata Formats</td>
<td>until (eg., YYYY-MM-DD):</td>
</tr>
<tr>
<td>List Sets</td>
<td>metadataPrefix:</td>
</tr>
<tr>
<td>List Identifiers</td>
<td>identifier:</td>
</tr>
<tr>
<td>List Records</td>
<td>set:</td>
</tr>
<tr>
<td>Get Record</td>
<td>resumptionToken:</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Language</th>
<th>Display</th>
<th>Schema Validation</th>
</tr>
</thead>
<tbody>
<tr>
<td>English</td>
<td>Parsed</td>
<td>None</td>
</tr>
<tr>
<td></td>
<td>Raw XML</td>
<td>Local mirror of schemata (Xerces)</td>
</tr>
<tr>
<td></td>
<td>Both</td>
<td>Online schemata (Xerces)</td>
</tr>
</tbody>
</table>

Send all comments to hussein@vt.edu --- Digital Library Research Laboratory@Virginia Tech
## 7.3. RE Browsing

### Archive Self-Description

<table>
<thead>
<tr>
<th>Repository Name</th>
<th>JCDL 2001 Picture Archive</th>
</tr>
</thead>
<tbody>
<tr>
<td>Base URL</td>
<td><a href="http://rocky.dlib.vt.edu/~jcdlpix/cgi-bin/OAI2.0/beta2/jcdl/oai.pl">http://rocky.dlib.vt.edu/~jcdlpix/cgi-bin/OAI2.0/beta2/jcdl/oai.pl</a></td>
</tr>
<tr>
<td>Protocol Version</td>
<td>2.0b2</td>
</tr>
<tr>
<td>Admin Email</td>
<td><a href="mailto:jcdlpix@rocky.dlib.vt.edu">jcdlpix@rocky.dlib.vt.edu</a></td>
</tr>
<tr>
<td>Earliest Datestamp</td>
<td>1970-01-01T00:00:00Z</td>
</tr>
<tr>
<td>Deleted Record Handling</td>
<td>no</td>
</tr>
<tr>
<td>Granularity</td>
<td>YYYY-MM-DD</td>
</tr>
</tbody>
</table>

### Other Information

```plaintext
description:
  toolkit:
  title: VTOAI Perl Data Provider
  author:
    name: Hussein Suleman
    email: hussein@vt.edu
  institution: Virginia Tech
  version: 3.04
  URL: http://www.dlib.vt.edu/projects/OAI/
```
7.4. RE Browsing

List of Sets

Click on the link to list the contents

JCDL Day Four

set description:
  dc:
    description: Pictures taken during JCDL Day Four

JCDL Banquet

set description:
  dc:
    description: Pictures taken during JCDL Banquet

JCDL Day Three
7.5. RE Browsing

List of Record Identifiers

Select a link to view more information

**header:**

identifier: oai:JCDLPICS:200105dle1

datestamp: 2001-06-27

setSpec: 200105dle

[display record in Dublin Core] [display metadata formats]

**header:**

identifier: oai:JCDLPICS:200105dle2

datestamp: 2001-06-27

setSpec: 200105dle

[display record in Dublin Core] [display metadata formats]
7.6. RE Browsing

List of Metadata Formats

Click on the link to view schema

<table>
<thead>
<tr>
<th>Prefix</th>
<th>Namespace</th>
<th>Schema</th>
</tr>
</thead>
<tbody>
<tr>
<td>dc2</td>
<td>[<a href="http://www.openarchives.org/OAI/2.0/oai_dc/">http://www.openarchives.org/OAI/2.0/oai_dc/</a>]</td>
<td>[<a href="http://www.openarchives.org/OAI/2.0/oai_dc.xsd">http://www.openarchives.org/OAI/2.0/oai_dc.xsd</a>]</td>
</tr>
<tr>
<td>oai_dc</td>
<td>[<a href="http://www.openarchives.org/OAI/2.0/oai_dc/">http://www.openarchives.org/OAI/2.0/oai_dc/</a>]</td>
<td>[<a href="http://www.openarchives.org/OAI/2.0/oai_dc.xsd">http://www.openarchives.org/OAI/2.0/oai_dc.xsd</a>]</td>
</tr>
</tbody>
</table>

[display record]
7.7. RE Browsing

List of Fields

header:
  identifier : oai:JCDLPICS:200105dle1
  datestamp : 2001-06-27
  setSpec : 200105dle

metadata:
  dc:
    title: 01dle1
    creator: Hussein Suleman
    subject: JCDL Day Four
    description: Jim French and Unmil Karadkar over lunch
    publisher: JCDL
    date: 2001-06-27
    type: image
    format: image/jpeg
    identifier: http://rocky.dlib.vt.edu/~jcdlpix/pictures/200105dle/01dle1.jpg
    language: en-us
    relation: http://www.jcdl.org
    rights: unrestricted
7.8. RE Multiple views of data

```xml
<?xml version="1.0" encoding="UTF-8"?>

  <responseDate>2002-05-26T19:59:35Z</responseDate>
  <request verb="GetRecord" metadataPrefix="oai_dc" identifier="oai:JCDLPICS:200105dle1" />
  <GetRecord>
    <record>
      <header>
        <identifier>oai:JCDLPICS:200105dle1</identifier>
        <datestamp>2001-06-27</datestamp>
        <setSpec>200105dle</setSpec>
      </header>
      <metadata>
        <oaidc:dc xmlns="http://purl.org/dc/elements/1.1/" xmlns:oaidc="">
          <title>01dle1</title>
          <creator>Hussein Suleman</creator>
          <subject>JCDL Day Four</subject>
        </oaidc:dc>
      </metadata>
    </record>
  </GetRecord>
</OAI-PMH>
```
7.9. RE Multilingual Support

Hier können Sie interaktiv die OAI-Kompatibilität Ihrer Archive verifizieren. [Bitte hier klicken um JavaScript zu verhindern]

Bemerkung: Um HTTP-Fehler zu vermeiden, warten Sie bitte bis eine Seite fertig geladen ist, bevor Sie klicken.

Bitte geben Sie die URL zu dem OAI Interface ein (alles vor dem "?") oder wählen sie ein vordefiniertes Tabelle.
7.10. RE Automatic Test Suite

Open Archives Initiative - Repository Explorer
exposer version - 1.44 : protocol version - 2.0b2 : May 2002

Open Archives Initiative :: Protocol for Metadata Harvesting v2.0b2
RE Protocol Tester 1.44 :: Virginia Tech DLRL :: May 2002

(1) Testing : Identify
URL : http://rocky.dlib.vt.edu/~jcdlpix/cgi-bin/OAI2.0/beta2/jcdl/oai.pl?verb=Identify
Test Result : OK
---- [ Repository Name = JCDL 2001 Picture Archive ]
---- [ Protocol Version = 2.0b2 ]
---- [ Base URL = http://rocky.dlib.vt.edu/~jcdlpix/cgi-bin/OAI2.0/beta2/jcdl/oai.pl ]
---- [ Admin Email = jcdlpix@rocky.dlib.vt.edu ]
---- [ Granularity = YYYY-MM-DD ]

(2) Testing : Identify (illegal_parameter)
URL : http://rocky.dlib.vt.edu/~jcdlpix/cgi-bin/OAI2.0/beta2/jcdl/oai.pl?verb=Identify&
Test Result : OK

(3) Testing : ListMetadataFormats
URL : http://rocky.dlib.vt.edu/~jcdlpix/cgi-bin/OAI2.0/beta2/jcdl/oai.pl?verb=ListMetadata
Test Result : OK
---- [ Sample Metadata Format = dc2 ]
### 7.11. RE Error in Response

**Archive Self-Description**

<table>
<thead>
<tr>
<th>Repository Name</th>
<th>Virginia Tech Electronic Thesis and Dissertation Collection</th>
</tr>
</thead>
<tbody>
<tr>
<td>Base URL</td>
<td><a href="http://oai.dlib.vt.edu:80/~hussein/cgi-bin/NDLTDErr1/VTETD.pl">http://oai.dlib.vt.edu:80/~hussein/cgi-bin/NDLTDErr1/VTETD.pl</a></td>
</tr>
<tr>
<td>Protocol Version</td>
<td>1.0</td>
</tr>
</tbody>
</table>

Error: Missing field: `<Identify>/adminEmail`  

**Other Information**

```yaml
description:
  oai-identifier:
    scheme: oai
    repositoryIdentifier: VTETD
    delimiter: :
    sampleIdentifier: oai:VTETD:etd-171110282978860

description:
  eprints:
    content:
      text: Theses and Dissertations produced by students at Virginia
    metadataPolicy:
      text: Metadata may be used by commercial and non-commercial users
    dataPolicy:
      text: Full texts are individually tagged and the rights statement
```
7.12. RE Error in XML

XSD Schema/Instance Validation Error!

Errors in XML instance

```xml
<?xml version='1.0'?>
<xsv docId='(http://www.openarchives.org/OAI/1.0/OAI_Identify)Identify' instanceId=''
<importURI='http://oai.dlib.vt.edu/OAI/1.0/OAI_Identify.xsd' namespace='http://www.openarchives.org/OAI/1.0/OAI_Identify.xsd' />
<importURI='http://oai.dlib.vt.edu/OAI/sprints.xsd' namespace='http://www.openarchives.org/OAI/1.0/OAI_Identify.xsd' />
<invalid char='4' code='cvc-complex-type.1.2.4' line='11' resource='file://tmp/file2VW'
<fsm>
  <node id='1'>
    <edge dest='2' label='(http://www.openarchives.org/OAI/1.0/OAI_Identify):responseData' />
  </node>
  <node id='2'>
    <edge dest='3' label='(http://www.openarchives.org/OAI/1.0/OAI_Identify):requestURL' />
  </node>
  <node id='3'>
    <edge dest='1' label='(http://www.openarchives.org/OAI/1.0/OAI_Identify):requestURL' />
  </node>
</fsm>
</xsv>
```
7.13. OAI Registry

The Open Archives Initiative

Registering as a Data Provider

Data providers who support the OAI protocol may choose to list their repository in the OAI registry. The goals of the registry are:

- Provide a publicly accessible list of OAI conformant repositories, making it easy for service providers to discover repositories from which metadata can be harvested.
- Provide a mechanism for data providers to ensure their conformance with the OAI protocol specification.
- Provide a means for the OAI to monitor use of the protocol and plan future activities and strategies.

This page allows you to register your repository by entering your BASE-URL in the text box at the bottom of this page. Before doing that, please read all of this instruction page so you understand what registration means and the choices you have.

Consequences of Registration
Protocol Testing
Conformance Testing
7.14. OAI Registry

The Open Archives Initiative

List of Registered, OAI-Conformant Repositories

<table>
<thead>
<tr>
<th>OAI Repository Identifier</th>
<th>Repository Name</th>
</tr>
</thead>
<tbody>
<tr>
<td>celebration</td>
<td>A Celebration of Women Writers</td>
</tr>
<tr>
<td>arlc</td>
<td>Alaska Native Language Center</td>
</tr>
<tr>
<td>arXiv</td>
<td>arXiv</td>
</tr>
<tr>
<td>CDLCIAAS</td>
<td>California International and Area Studies Digital Repository</td>
</tr>
<tr>
<td>NUDL-CSTR</td>
<td>NUDL-CSTR</td>
</tr>
</tbody>
</table>

This application allows you to browse the current list of OAI conforming repositories. Currently there are 29 such repositories. The table may be sorted either by the OAI Repository Identifier or by the Repository Name.

You may retrieve information about an OAI repository by selecting one of the rows in the following table. You may view the registration record from the database; alternatively, if your browser can render XML, you may issue the Identify request to the selected repository and receive the current XML response.

Sort repositories by:
- Repository Name
- OAI identifier

- view registration record
- issue Identify request
Validator for XML Schema 20000922 version, XML Output

XSV version: XSV 1.176/1.87 of 2001/02/16 16:38:43

NOTICE: This is an ALPHA TEST of a service for a work-in-progress specification. This version is for schema documents with the namespace URI http://www.w3.org/2000/10/XMLSchema and is being actively developed: see XSV for XML Schema 200004007 version for the no longer maintained previous version, for schema documents with the namespace URI http://www.w3.org/1999/XMLSchema.

Use this form for checking a schema which is accessible via the Web, and/or schema-validating an instance with a schema of your own.

Address(es):

JCDL 2002
8. Service Providers

- How to Harvest
- Policies
- Intermediate systems
- Tools
- Case Study: ARC
- Case Study: NDLTD
8.1. How To Harvest

- Identify to get basic information
- ListIdentifiers, followed by ListMetadataFormats for each record and then GetRecord for each id/metadata combination
  - No. of short HTTP requests = 1+n+n \times m
    - n=no. of identifiers, m=no. of metadata formats
- ListRecords for each metadata format required
  - No. of long HTTP requests = m
    - m=no. of metadata formats
8.2. Policies

- Use schedule for harvesting regularly
- Store date when last harvested (before you start)
- Use a two day overlap (or one day if your archive uses proper UTC datestamps)
  - New items may be added for the current day
  - Timezones create up to a day of lag if you ignore them
  - If the source uses correct UTC datestamps and second granularity then only 1 second of overlap is needed!
- Each time a record is encountered, erase previous instances
8.3. Intermediate Systems

- Both a data provider and service provider
- All harvested data must have the datestamps updated to the date on which the harvesting was done
- Identifiers retain their original values
- Note: Consistency in the source archive propagates, but so does inconsistency!
8.4. Tools

- Check OAI website for sample code
- XML parsers – depending on platform – check W3C
- XML Schema validators
  - Very few available – the reference version works but may not be easy to install
  - Ignore validation if you can trust the source
- Sample data providers – check the OAI website for a list of conformant public archives
8.5. Case Study: ARC

Cross Archive Searching Service

This is page 1, hits (1–10) of total 66 hits.

Results Pages: 1 2 3 4 5 6 7

SEARCH RESULTS

Title: Chemical Encoding for Information Visualization: Using Icon Color, Shape, and Size to Convey Nominal and Quantitative Data

Creators: Nowell, Lucille Terry

Description: Producing a user interface design to visualize search results for a digital library called En

Archive: NDLTD

Document ID: etd-111897-163723

This prototype is based on the UPS project and the NCSTRL+ based digital library developed by Old Dominion University.
8.6. Case Study: NDLTD

Search/Browse Engines
- VTLS Virtua
- MARIAN

Other Services
- Recommender
- Cross-Ref.

NDLTD ETD Union Catalog

Institutions
- Virginia Tech
- Humboldt U.
- U. Oldenberg

...
9. OAI Communities

- Shared Metadata Formats
- Shared semantics
- Layering over OAI
- Closed OAI networks
- OAI within the DL
9.1. Shared Metadata Formats

Use metadata formats accepted within a community to convey more specific information

Examples

- E-Print format (under development)
- ETD-MS for theses and dissertations
- VRA Core for multimedia
- IMS Metadata for educational material
9.2. Shared Semantics

- Develop a shared understanding for the meanings of fields

Examples

- Developing controlled vocabularies for fields
- Using specific fields for external links (OAI recommends using identifier in DC for this)
- Choosing from among existing standards (like language names)
9.3. Layering over OAI

- Convert OAI records into more standard formats like MARC communications format
- Collapse multiple requests into one to make harvesting easier
- Name authority system (developed at OCLC) piggybacks name resolution over the OAI protocol
9.4. Closed OAI networks

- Data providers need not go public!
- Within an organization, OAI can be used for data transfer among heterogeneous systems
- More control over use, making global optimizations possible (like harvesting schedules and choice of metadata formats)
9.5. OAI within the DL

Use the OAI protocol as the basis for components to communicate

Examples

- Search Engines could use dynamic sets to correspond to search results
- Browsing can be directed by sets
- Reviews and Annotations can each be independent OAI data providers

Open Digital Libraries project to investigate this approach:

- http://oai.dlib.vt.edu/odl
10. Now What?

- 1-2-3 Recipe
- Future of Metadata Harvesting Protocol
- Future of OAI
- Links
10.1. 1-2-3 Recipe

-do I really want to do this?
-do I have an accessible metadata source?
-do I have a server to host the OAI script/program?
-can I satisfy the requirements to be a data provider?
-can I write the code or modify a template or hire a programmer to do either?
10.2. Future of Protocol

Version 2.0
- Released ~ June 2001

SOAP version
- Envisioned for near future
10.3. Future of OAI

- Advocacy for easier access to information
- New protocols/tools to support this mission
- Research projects to test theory underlying current architecture e.g. Cyclades
10.4. Links

- Open Archives Initiative
  - http://www.openarchives.org

- OAI Metadata Harvesting Protocol
  - http://www.openarchives.org/OAI/openarchivesprotocol.htm

- Virginia Tech DLRL OAI Projects
  - http://www.dlib.vt.edu/projects/OAI/

- Repository Explorer
  - http://purl.org/net/oai_explorer

- NDLTD
  - http://www.ndltd.org
10.5. More Links

- ARC Cross-Archive Search Service
  - http://arc.cs.odu.edu/
- XML Schema Validator
  - http://www.w3.org/2001/03/webdata/xsv
- Dublin Core Metadata Initiative
  - http://www.dublincore.org
- E-Prints DL-in-a-box
  - http://www.eprints.org
- XML Tools at W3C
  - http://www.w3.org/XML/#software
That’s All Folks!